15th International Roundtable on Business Survey Frames
Washington, D.C. – October 22 – 26, 2001

Session No 6
Paper No 4
Bill Iwig, National Agricultural Statistics Service, USDA

Use of Federal Tax Information in Building the 2002 Census of Agriculture Mail List

1. Introduction

The Census of Agriculture Act of 1997 transferred responsibility for conducting the Census of Agriculture from the Bureau of the Census, Department of Commerce, to the National Agricultural Statistics Service, Department of Agriculture (NASS/USDA). Subsequently, the U.S. Code was amended (Title 26, Sec. 6103(j)(5)) to permit the Internal Revenue Service (IRS) to provide specific federal tax information (FTI) items to NASS/USDA for use in conducting the Census. Previously, IRS had provided these FTI items to the Bureau of the Census. The FTI items include the names and addresses of individuals, partnerships, and corporations that report farm income or have been assigned an agricultural activity code for their business. Prior to this regulation change, NASS had not been permitted to use FTI for supporting its statistical program. But a major requirement for conducting the Census is to build an efficient mail list that provides maximum coverage of the farm operations in the United States. This law was changed because of the critical importance of FTI to building the mail list and conducting an accurate Census of Agriculture.

Even though the IRS is authorized to provide FTI to NASS for census purposes, it will do so only if strict confidentiality safeguards are in place. Confidentiality is a sensitive issue in the U.S. and must be maintained at the highest level possible. As noted by the IRS, "in fostering our system of taxation the public must have and maintain a high degree of confidence that the personal and financial information furnished to us is protected against unauthorized use, inspection, or disclosure." (Internal Revenue Service, 1999)

In response to these requirements, NASS developed security procedures for processing the FTI to maintain strict confidentiality. These procedures limit the number of NASS employees with access to the FTI, include strict physical security requirements to protect the data, and avoid adding any actual FTI to the NASS farm register. The purpose of the FTI is to identify potential farms or agricultural operators from third-party sources that should be added to the farm register. The actual 2002 Census mail list will be extracted from the farm register and will include all records identified as "active" farms and potential farms. A consequence of these procedures is that Census coverage may not be quite as high as if all agricultural tax return records were actually added to the register for census enumeration. NASS will be monitoring this issue as the mail list development proceeds.

This paper will document the need for using FTI data in building the Census mail list, describe the actual FTI requested by NASS, more fully describe the IRS recommendations for protecting the confidentiality of the FTI, document the NASS plan for processing the FTI, and provide preliminary results on the impact of the FTI in building the 2002 Census population frame or mail list.

2. The Need for Federal Tax Information (FTI) for Building the Census Mail List

A major requirement for conducting any census is to build a population frame that is as complete as possible. For the Census of Agriculture, the population frame should cover all farms. A farm is defined by USDA as any establishment from which $1,000 or more of agricultural products were sold or would normally be sold during the year. FTI is one of the most complete sources of farm operators in the U.S. and, consequently, is an important source for building the Census of Agriculture mail list. Name lists from other government agencies, farm organizations and producer groups are also used. Table 1 contains counts that indicate the impact of the FTI from tax years 1995 and 1996 on the 1997 Census mail list. Sources used to build the mail list included FTI from 1995 and 1996, the 1992 mail list, the 1996 NASS farm register, and other special lists from commodity organizations. The table shows that the FTI was at least one source for 2.44 million records of the 3.22 million records on the mail list. Also, the table shows that FTI was the only source for .24 million farms out of 1.68 million

actual farm reports. So without the FTI, the Census would have missed approximately one-seventh of the farms actually counted.

Table 1: 1997 Census of Agriculture Mail List Counts (Millions)

Source	Total	In-Scope (Farm)	Out-of-Scope (Non-Farm)	Non-Respondent	Not Delivered and Other
Total	3.22	1.68	.99	.43	.12
Any FTI	2.44	1.43	.65	.32	.04
FTI Only	.60	.24	.23	.11	.02
No FTI	.78	.25	.34	.11	.08

But the FTI is not a perfect or complete source either. There were .23 million records on the 1997 mail list that had FTI as their only source, but their operations didn't have $1,000 in agricultural sales and were not classified as farms. Also, there were another .25 million farms represented by mail list records that were not included on the 1995 and 1996 FTI files. In addition, an independent national area frame survey was used to measure the coverage of the Census mail list. The names of all farms operating land within the randomly selected area frame land segments were checked against the mail list to determine if the farm was included on the mail list or not. The 1997 area frame survey estimated that approximately .30 million farms were in operation in 1997 but not on the mail list. So, in total, there were approximately .55 million farms in 1997 that were not represented by a federal tax record. Some of these were extremely small farm operations with little or no actual agricultural sales who did not report any agricultural income on their tax form. Others possibly filed their tax forms under different names than reported on the Census, so they could not be linked to FTI data.

3. Description of the Federal Tax Information (FTI) Requested by NASS from IRS

All individuals who receive taxable income in the U.S. are required to complete a *U.S. Individual Income Tax Return Form 1040* each year. One of the items on the form is "Farm income or (loss)." All individuals who report farm income or loss are then required to complete a *Profit or Loss From Farming Form 1040F*. NASS has requested and is receiving data from IRS for all 1040F forms, including:

1. Taxpayer and spouse social security numbers (SSNs)
2. Taxpayer name
3. Mailing address (Street, City, State, Zip Code)
4. Principal business activity (PBA) code
5. Employer identification code (EIN)
6. Gross income.

A Social Security Number (SSN) is assigned by the Social Security Administration to U.S. citizens and legal aliens who apply for a number in order to receive Social Security benefits. Employer Identification Numbers (EIN) are assigned by the Internal Revenue Service to all legal corporations and partnerships for Federal tax reporting purposes. Sole proprietors who hire workers and file Social Security tax payments are also required to have an EIN. Gross income is used for prioritizing processing of the FTI records. Other administrative items from the 1040F records are included on the FTI data files as well as two additional data items (Taxable cooperative distributions on an accrual basis and Income from custom hire and machine work on a cash basis) that will be evaluated for possible use in screening out non-farm records from the data file.

Another tax return form of interest to NASS is the *Employer's Annual Tax Return for Agricultural Employees Form 943* used to report income tax withheld and social security and medical insurance (Medicare) taxes on wages paid to farmworkers by individuals, partnerships, or businesses. NASS is receiving data from all Form 943 records, including:

1. Individual or business name
2. Mailing address (Street, City, State, Zip Code)
3. Employer Identification Number (EIN)
4. Wages subject to Medicare taxes.

All partnerships and incorporated businesses that receive taxable income are also required to complete tax returns. Partnerships use the *U.S. Partnership Return of Income Form 1065* and incorporated businesses use the *U.S. Corporation Income Tax Return Form 1120*. NASS receives selected identification and data items from all Form 1065 and Form 1120 records that:

1. Have an associated Form 943, indicating an agricultural operation
2. Have a non-zero amount for Net Farm Profit (Loss)
3. Have a Principal Business Activity code in the agricultural range.

There can be numerous combinations of FTI records associated with a single farming operation. Following are a couple common situations.

1) Any agricultural partnership should have a Form 1065 record for the partnership name and associated 1040F records for each partner, reporting their partnership share of income and expenses. If the partnership employs agricultural workers, there should be a Form 943 with the partnership name as well.

2) Any incorporated farming operation should have a Form 1120. If the business employs agricultural workers, there should be a Form 943 with the business name as well.

3) A farming operation can file multiple forms within a year or over years under different names.

This apparent duplication should be identified and eliminated in the record linkage procedures for processing the data. But, unfortunately, record linkage is not a perfect process, so some duplication undoubtedly will result.

In 1997, Table 1 shows that FTI was at least one source for 2.44 million records on the mail list. Of those, 2.19 million had an associated 1040F form and .51 million had at least one of the other forms (943, 1065, or 1120) as a source. While the 1040F is the dominant source of names, the other forms are also a significant source of names for the Census.

4. The IRS Confidentiality Safeguard Guidelines

The IRS publication "Tax Information Security Guidelines for Federal, State, and Local Agencies - Safeguards for Protecting Federal Tax Returns and Return Information" documents the FTI security guidelines recommended by IRS. FTI data has been categorized by the IRS as a high security item, meaning that it requires "greater than normal security due to their sensitivity and/or the potential impact of their loss or disclosure." In general, any agency receiving FTI must be able to show, to the satisfaction of IRS, their ability to protect the confidentiality of the data from any unauthorized use or disclosure. Also, FTI can only be used for the specific purpose included in the written request to IRS for the data.

Protecting the confidentiality of the FTI involves providing physical security of the data and restricting access to the data in order to protect against unauthorized use or disclosure. Providing physical security of the data can involve locked containers, vaults, locked rooms, restricted areas, guards, and electronic security systems. It also involves developing secure hardware, software, and data transmission provisions. In addition, the receiving agency must maintain an audit trail of all access to the tax data and of all operations, procedures, or events occurring on the system. Each agency has the responsibility to develop their own physical security plan, considering their specific setting, facility, and equipment.

In addition to providing strict physical security of the data, IRS also recommends that agencies restrict access to FTI to specific users for specific items. In general, IRS recommends "that access to FTI

must be strictly on a need-to-know basis." The data can not be indiscriminately circulated within the agency. Looking up data without a purpose is not permitted. Also, agencies should take steps to avoid commingling FTI with other data. If FTI data are merged with other data to create new data sets, those data sets must be protected with the same safeguards as the original FTI data sets. IRS recommends that commingling be avoided if at all possible.

IRS also recommends that Employee Awareness programs be implemented promoting the agency's safeguard measures to protect the confidentiality of the FTI. Employees should be certified "that they understand security procedures and instructions" and also understand the penalties for unauthorized disclosure or viewing of FTI.

Finally, all confidentiality safeguard plans must be approved by the IRS. Each agency is required to submit a Safeguard Procedures Report, documenting how the FTI is processed and their procedures for protecting against unauthorized use. This report must be approved before any FTI are sent to the requesting agency. The report is then updated at least once a year, documenting any changes in procedures and certifying that the FTI has been protected during the previous year.

5. The NASS Plan for Processing Federal Tax Information

The NASS plan for processing the Federal Tax Information (FTI) received from IRS was originally documented in the January 2000 Safeguard Procedures Report submitted by NASS to IRS. The Report was updated in January 2001. This report documents plans for handling the FTI data tapes, processing procedures for protecting the confidentiality of the actual FTI data, and other security measures.

The FTI data tapes are initially sent to the NASS Security Officer who then delivers them to the Information and Technology Division for the actual processing. NASS started receiving tapes in September 2000 and receives tapes containing new data each week. An FTI Data Tape Transmittal Form is used to document the receipt and use of each of the tapes at each step of the process.

A secure UNIX system is used for all processing of the data with access limited to specific individuals named in the Safeguards Report. Access restrictions are ensured by only activating user IDs for the identified individuals. These individuals can then access the FTI at their own workstations through an independent secure firewall. To help ensure that the data are used only in an "official" manner, each and every system and database activity is audited. Audit records are retained for six years. Any user who accesses the system through the local console in the limited access area is video taped and a record of the access is maintained. Also, all individuals with access to the data at their own workstations sign a form pledging to keep the data confidential under penalty of the law.

The contents of each FTI data tape are loaded into a Sybase database on the secure UNIX system. Two backup tapes of the database and two backup tapes of the UNIX system are then created. One set of tapes is kept in a secure, locked cabinet on-site and the other set is kept off-site. These backup tapes are kept for six years. A tracking form is created to audit these tapes until destruction. The original FTI data tapes are then returned to the IRS.

The FTI data in the Sybase database are then run through the NASS record linkage system to identify farm operations represented by the FTI data that are not already on the NASS farm register as an active record. An active record represents an operation currently believed to be in business. A key feature of this process is that actual FTI are never added to the farm register, thus protecting the confidentiality of the FTI and reducing the audit requirements of the farm register. Rather, the FTI data records are used only to identify potential farm operations on other third-party sources. The goal is to have third-party sources that, in combination, provide nearly complete coverage of the farm population. But typically, these sources contain many names that are not associated with farm operations. So the FTI data are used only to identify names on these sources that have a high likelihood of representing a farm and, consequently, to eliminate the non-farm records.

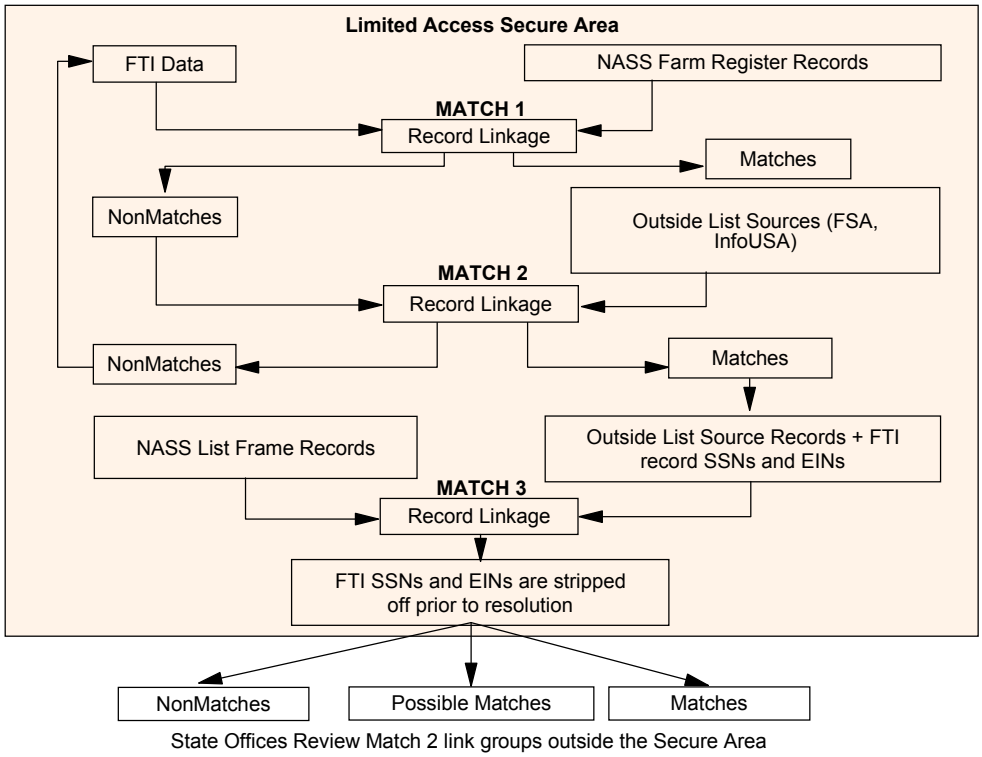
Currently, NASS is using two third-party sources. The first is a data file from the U.S. Department of Agriculture's Farm Service Agency (FSA) that contains the names of all participants in the various USDA farm support programs. Many family members and landlords may receive farm program payments and be included on the FSA list but they do not actually operate a farm. Consequently, this

data file contains over 11 million records compared to the NASS estimate of 2.17 million farms in the U.S. in 2000. The second third-party source is the commercially available InfoUSA database which contains name and address information for approximately 120 million consumer households in the U.S.. The database is compiled and continually updated from many different sources including telephone directories, automobile registrations, new births, credit information, and real estate transactions. Addresses are kept current using the U.S. Postal Service National Change of Address (NCOA) system. National statistics show that approximately 20 percent of all consumers move once a year. But the continual updating conducted by InfoUSA helps maintain an accuracy rate of about 90 percent.

As FTI records are matched to NASS farm register records during record linkage, the FTI sequential identification number, the NASS farm register identification number and the NASS active status code are posted to a cross reference table. This table is used for analysis of the record linkage procedures.

Figure 1 contains a flow diagram of the record linkage procedures. A series of steps are involved as described below.

Figure 1: Flow Diagram of Record Linkage Processing



Match 1: All FTI records are matched to all NASS farm register records

The first probabilistic record linkage process will match all newly received FTI data records and FTI records not previously matched to a NASS record or outside source record to all records on the NASS farm register. The parameters for this match are designed such that each FTI data record will be classified either as a definite match or nonmatch to an existing NASS farm register record. There will be no staff review of match results from this match. Variables which will be included in the match are items such as SSN, EIN, and various combinations of standardized and reformatted names and addresses.

FTI records which match existing NASS farm register records will not be passed to the subsequent record linkage processes. Their linkage information will be updated on the cross reference table,

described above. For some cases where an FTI record matches a NASS farm register record that has an ineligible active status code for the Census of Agriculture, the status code of the NASS record is updated to a value which is included in the Census of Agriculture. An exception is if the FTI record matches a NASS record coded as deceased. In that case, the FTI record is treated as a non-match.

FTI data records which do not match an existing NASS farm register record are passed on to the subsequent record linkage processes.

Match 2: All nonmatched FTI records are matched to records from outside list sources

The second record linkage process will compare FTI data records, which were not linked to NASS farm register records in the first match, to records from a third party list source. Matching will compare items such as SSN, EIN, and various combinations of standardized and reformatted names and addresses. Record linkage cutoff boundaries will be set such that each record pair will either be classified as a definite match or a nonmatch, based on the likelihood that the records represent the same operation. There will be no staff review of match results at this stage.

FTI records that do not match any record from the third party list source will not go on to the next match process. Rather they will be appended to each succeeding batch of incoming FTI records and reprocessed through the match cycle until they are linked either to a NASS farm register record or a third party source record.

When an FTI record matches a record from a third party list source, the corresponding record from the third party list source, the FTI SSN, the FTI EIN, and the FTI record's sequential identification number are written to a file that is used for the third record linkage process. The FTI-SSN and FTI-EIN are only used for record linkage purposes. They never leave the secure area.

Match 3: All third party source records which matched FTI records in Match 2 are matched to all NASS farm register records

The third record linkage process will match the third party "matched records", including the FTI records' SSN and EIN, to all records currently on the NASS farm register. Unlike the previous matches, the cutoff boundaries for this match are set such that each record pair is classified as either a definite match, possible match, or a nonmatch. No possible matches were identified in the previous matches.

After the groups of linked records are formed, the FTI-SSN and FTI-EIN are removed from the output records. If a third party source record has a SSN or EIN, the SSN and/or EIN are retained on the record. The NASS generated sequential FTI identification number is also retained on the records for tracking purposes. The output records are then transferred out of the secure area for staff review. Only third party source name, address, and identification information and the FTI sequential ID are contained on the output record. No FTI data are included on any of the output records.

The output records are populated into a Sybase database outside of the secure area for staff review. State Office personnel review the possible matches and a subset of the definite matches and the nonmatches using Windows-based review screens. The review procedure is one that is familiar to NASS employees who process new list sources on an ongoing basis. (See the USA/NASS paper presented at the 13th Roundtable, "Record Linkage at NASS Using Automatch", for more details.)

Once resolution is complete, each third party source record is either classified as a match or a nonmatch. Name, address and phone number information for matching records may be updated on the NASS farm register if more current information is present on the third party list source record. When appropriate, the active status code of the NASS farm register record may be updated so that the record will be included in the Census of Agriculture. Third party list source records that do not match NASS farm register records will be added to the NASS farm register as inactive records. The final linkage information, including the NASS farm register ID for matching records, is transferred back to the secure area in order to update the cross reference table.

In all cases, the federal tax filer name and address as well as all other FTI data are only used during record linkage in the limited access secure area. FTI data are not posted or commingled on the NASS

farm register. The only records that are added to the NASS farm register are third party source records that match an FTI record and do not match an existing NASS record.

Current plans are to retain nonmatching FTI records in the limited access secure area and combine them with future receipts of FTI data until they can be linked to a NASS farm register record or a third party list source. If a significant number of FTI records never match third party source records, their exclusion from the NASS farm register may hurt Census coverage. This situation will be monitored during the list-building process.

6. Preliminary Results of FTI Processing

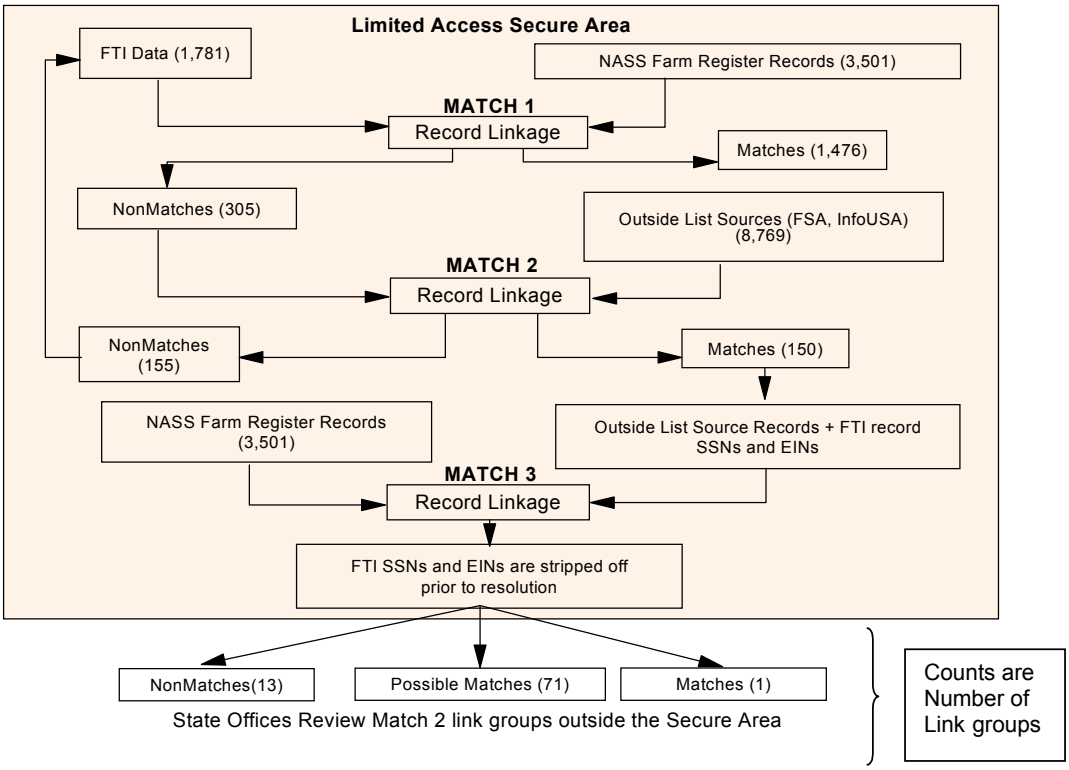
NASS began receiving FTI tapes containing 1999 and 2000 tax data from IRS in September 2000 and started the Batch 1 record linkage processing described above in April 2001, on a state by state basis. Subsequent batches of FTI tapes will be processed on a continual basis.

A major goal for NASS is to minimize the differences between the 2002 survey estimates and the 2002 Census results for common items, such as acres of corn or total cattle. In order to do this, the sampling populations for the various surveys need to be as similar as possible to the Census mail list. Consequently, NASS intends to process as many of the FTI records as possible during 2001 to identify FSA and InfoUSA records that have a high liklihood of being a farm for the 2002 Census. These potential farm records would then be contacted to collect basic agricultural data maintained on the farm register as “control data” in order to activate them for the 2002 survey program. New farm register records are initially added as inactive records until each operation is contacted to verify that the operation qualifies as a farm and to obtain control data.

For Batch 1, specified state-level gross income or receipts cutoffs were used to identify larger records for processing and data collection during 2001. All FTI records with values below these cutoffs are being held for processing in 2002. NASS decided to use this approach in order to target our limited budget and human resources during 2001 on the larger records which will have more of an impact on insuring the 2002 survey and Census results are comparable. The smaller records have more impact on Census farm counts of various categories. These records will be run through record linkage in 2002 to identify potential farm records on third party sources for inclusion on the Census mail list.

Figure 2 shows the results of the Batch 1 processing. The only third-party list source for Batch 1 was FSA data. The InfoUSA data file was not yet available at the time of Batch 1 processing. Only 44 of the 50 states were processed, excluding Arizona, Colorado, Indiana, Maryland, New York, and Texas. All counts shown in Figure 2 exclude these six states.

Figure 2: Batch 1 Processing Record Counts (000)



As indicated, approximately 1,780,000 FTI records were processed in Batch 1. The total number of FTI records available was about 3,880,000. About 2,100,000 smaller records are being held until 2002 for processing. The NASS farm register contains about 3,500,000 active and inactive records. Currently, 1,800,000 are categorized as active and 1,700,000 are inactive.

The Match 1 record linkage identified 305,000 FTI records that did not match the current NASS farm register. The parameters for this record linkage processing were set conservatively so that FTI records determined to match farm register records by the system had a high probability of being true matches. These 305,000 records represent potential farming operations that are currently not on the NASS farm register. But, as documented previously, NASS will not add these records to the NASS farm register. Instead, we will use these records to identify records on other third-party sources that should be added to the NASS farm register.

In Match 2, these non-matching FTI records were matched against the FSA data file, containing approximately 8,769,000 records over the 44 states. Again, parameters were set to identify high likelihood matches. Only about 150,000 records matched. The other 155,000 are being held for Batch 2 processing, which will also use the InfoUSA data file as a third-party source. So, it is at this step that we have used the FTI data to identify FSA records that are not currently represented on the NASS farm register and should be added as a potential farm. For each matching pair of FTI and FSA records, the FTI Social Security Number (SSN) and Employer Identification Number (EIN) are added to the FSA record for use in the next match. It should be noted that, for matched records, the FSA name and address may not match exactly the FTI name and address. But, based on the record linkage results, we assume the FSA name and address represents the same farm as the FTI name and address. Following are a couple examples of matching FTI and FSA names and addresses to illustrate possible differences. These matches are heavily dependent on the matching Employer Identification Number (EIN) or Social Security Number (SSN) for each pair.

Example 1: FTI name and address

Larry & Pat Hollis Farms
Hollis Larry & Pat Gen Ptrs
124 Bolger Ln
Churchville AR 70000
EIN: 111223333

FSA name and address

L & P Farms Ptr
RR 1 Box 99
Churchville AR 70000
EIN: 111223333

Example 2: FTI name and address

Terry L & Mary J Gish
PO Box 888
Oldtown IL 60000
SSN: 333445555

FSA name and address

Terry Gish
Box 888
Oldtown IL 60000
SSN: 333445555

The FTI records that matched the 150,000 FSA records were originally determined to be non-matches to the NASS farm register. However, at this point, we match the FSA records back against the NASS farm register to determine if any of the FSA records now match the register, due to the type of name differences noted above. The results show that about 1,000 FSA records actually do match the register, based on the record linkage output. The record linkage parameters for Match 3 are designed to create link groups containing definite matches, definite non-matches, and possible matches. Multiple FSA records can occur in the possible match and definite match link groups. The results indicate that, on average, these link groups contain two FSA records, reflecting that multiple FTI records were originally provided for many farm records. These link groups are sent to our State Statistical Offices for manual resolution. The FTI SSN and EIN are stripped from the records before being sent, again to protect the confidentiality of the FTI data. About 49,000 potential farm records were added to the farm register from Batch 1. In addition, another 19,000 previously inactive records were assigned a status code to be included in the Census of Agriculture mail list.

The 155,000 FTI records that are non-matches to the third-party FSA data source out of Match 2 are a big concern to NASS. These records represent potential farms that are not being added to the farm register in order to protect the confidentiality of the FTI data. Preliminary results from Batch 2 processing, which uses the InfoUSA database as a third-party source, indicate that approximately 55% of these will match InfoUSA and, consequently, can be added to the farm register. A preliminary review of the remaining non-matching records indicate that many of these records are linked to each other. They may represent records from multiple years for the same farm or they may refer to a partnership or corporate farm that provides multiple original FTI records. Others may be represented on the farm register under a different name that could not be matched by the record linkage system. NASS will continue to monitor these results.

7. Summary

NASS is receiving federal tax information (FTI) data from the Internal Revenue Service for use in conducting the Census of Agriculture. In order to receive this data, NASS developed strict security and confidentiality procedures to protect the data. These procedures limit the number of NASS employees with access to the FTI, include strict physical security requirements to protect the data, and

avoid adding any actual FTI to the NASS farm register. The FTI is only used to identify potential farms or agricultural operators from third-party sources that should be added to the farm register for the Census mail list. Consequently, not all of the FTI records that do not match the current NASS farm register are added to the register. NASS is monitoring the impact of these procedures on expected coverage of the Census mail list.

8. References

Internal Revenue Service (1999), "Tax Information Security Guidelines for Federal, State, and Local Agencies." Publication 1075 (Rev 3-99), Washington, D.C.